ExHuBERT: Enhancing HuBERT Through Block Extension and Fine-Tuning on 37 Emotion Datasets

Shahin Amiriparian[1], *Filip Packań*[2], *Maurice Gerczuk*[2], *Björn W. Schuller*[1,2,3]

[1]CHI – Chair of Health Informatics, MRI, TU Munich, Germany
[2]Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College, UK

shahin.amiriparian@tum.de

Abstract

Foundation models have shown great promise in speech emotion recognition (SER) by leveraging their pre-trained representations to capture emotion patterns in speech signals. To further enhance SER performance across various languages and domains, we propose a novel twofold approach. First, we gather EmoSet++, a comprehensive multi-lingual, multi-cultural speech emotion corpus with 37 datasets, 150,907 samples, and a total duration of 119.5 hours. Second, we introduce ExHuBERT, an enhanced version of HuBERT achieved by backbone extension and fine-tuning on EMOSET++. We duplicate each encoder layer and its weights, then freeze the first duplicate, integrating an extra zero-initialized linear layer and skip connections to preserve functionality and ensure its adaptability for subsequent fine-tuning. Our evaluation on unseen datasets shows the efficacy of ExHuBERT, setting a new benchmark for various SER tasks. Model and details on EMOSET++: https://huggingface.co/amiriparian/ExHuBERT.

Index Terms: affective computing, speech emotion recognition, transformers, deep learning

1. Introduction

Speech Emotion Recognition (SER) has a rich research history going back to the 1970s (first patents) [1] and 1990s (first research papers) [2] and while it has reaped the benefits of deep learning, a core issue remains to this day: While there are many available databases of emotional speech, most of them only contain comparatively few samples or speakers, hindering effective single-corpus training of large neural networks [3]. As a response to this circumstance, cross- and multi-corpus SER has established itself as a highly important research direction [4]. Due to the fact that databases in the field often differ significantly in recording settings, nature of speech and emotion portrayal (acted, elicited, natural), language, and other factors, successful approaches have employed special strategies and architectural considerations such as domain adaptation [5] or adapter transfer learning [6] to achieve satisfactory performance. Other efforts have gone towards making deep learning models more robust against distortions, noise, and other variations in the speech signal [7, 8].

However, the recent paradigm shift in general deep learning towards large, Transformer-based models has already impacted the field [9]. The Transformer architecture's inherent capability of learning arbitrary structural information from high-dimensional data combined with the exploitation of huge amounts of unlabeled data through self- and unsupervised learning has significantly reduced the need for human-annotated corpora for training powerful and transferable models [10, 11]. Specifically for the fields of speech recognition and analysis, pre-trained Transformers such as Wav2Vec2.0 (W2V2) [12] or HuBERT [13] have shown considerable generalization capabilities. By exploiting self-supervision on large amounts of unlabeled speech data, these models learn to effectively capture the structure of spoken language. For a wide range of downstream tasks, pre-trained transformers provide competitive performance as feature extractors [14, 15] or through finetuning, e. g., for SER and speaker identification [16]. Wagner *et al.* [9] finetune wav2vec and HuBERT models on MSP-Podcast [17] and show that their best models provide state-of-the-art performance on a number of SER corpora. They further trace the models' efficacy to a number of beneficial characteristics induced by both pre-training and the transformer architecture itself, such as implicit modeling of linguistic information and a general resilience against speaker, gender, or domain variations.

While these and other works have convincingly made the argument for large, pre-trained Transformer models in SER, none have investigated whether their generalisability and robustness can enable effective learning of a single, transferable model on a heterogenous set of databases without the need for domain or corpus adaptation strategies. In the present study, we aim to fill this gap by evaluating the capability of large audio transformers to learn salient features for SER by multi-corpus finetuning. For this purpose, we build on the work of [6], introducing EmoSet++, integrating 37 SER corpora spanning 15 languages. We then fine-tune HuBERT on the assembled corpus and compare its transfer learning performance to other large pre-trained models on 6 additional SER databases. Finally, we introduce ExHuBERT, which integrates EMOSET++ finetuning with Backbone Block Expansion (BBE) – a technique recently introduced in LLAMA Pro [18] – to deliver a state-of-the-art model and training strategy for emotion recognition.

2. EmoSet++

We introduce EMOSET++, a comprehensive multi-lingual, multi-cultural speech emotion corpus. It extends EmoSet [6] and integrates 37 unique emotion datasets with 150,907 speech recordings and a cumulative length of 119.5 hours. Most of the datasets that we have included comprise common languages such as English, German, or Mandarin, alongside rarer ones like Persian or Urdu. For all corpora, we create speaker-independent splits. To facilitate training, all distinct dataset labels (comprising 106 emotion classes) are mapped to six classes, representing combinations of low/high arousal and negative/neutral/positive valence. The mapping is based on Russel's Circumplex of Affect [19]. The dataset splits were derived through three methods: adopting from the collected dataset, manual construction for speaker independence, or simply dividing it into 10 % partitions for both testing and validation in cases where speaker

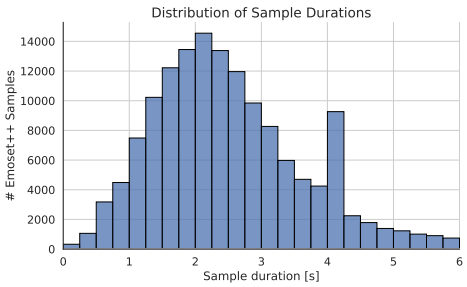

Figure 1: *Distribution of speech sample durations within* EMOSET++. *Predominantly, the samples fall within the range of* 0 *to* 5 *seconds in length.*

information was not explicitly provided. We do not conduct any re-weighting or re-balancing of speech samples for our machine learning experiments. Figure 1 displays sample durations ranging from 0 to 6 seconds, with longer samples excluded for readability. Most samples fall in $[0-5]$ second range.

3. Proposed Approach

We propose an enhanced version of HuBERT [13] through fine-tuning on EMOSET++ and incorporating Backbone Block Expansion (BBE), denoted as ExHuBERT. Specifically, we fine-tune the encoder part of the HuBERT architecture on 37 diverse emotion datasets, which span various languages and cultural backgrounds (cf. Section 3.1). After the weights are updated, we conduct backbone extension (cf. Section 3.2), and in the final step, we evaluate the performance of the backbone extended HuBERT (ExHuBERT) on 6 unseen test emotion datasets (cf. Section 4).

3.1. Fine-Tuning HuBERT on EmoSet++

Our system is built upon the Transformer architecture HuBERT [13], which has shown promising results in SER [9, 28, 29]. A simple linear layer is added on top for the classification of the 6 mapped arousal valence classes. Fine-tuning is conducted in a round-robin fashion, ensuring each dataset contributes equally to the model. Upon achieving a state where our model spans multiple domains and languages, we utilize its transfer learning and generalization capabilities within ExHuBERT.

3.2. Backbone Block Expansion

After the fine-tuning process, we duplicate each encoder layer along with its weights. This augmentation results in an expanded version of HuBERT (ExHuBERT), featuring a total of 48 layers. The newly added layers are inserted after the original ones, accompanied by a skip connection to maintain the original layer's behavior. To stabilize the training process after layer duplication, we add a Zero Linear Layer (ZLL) at the end of each duplicated layer. The ZLL comprises initialized zero weights, ensuring that the output of the copied layers initiates from zero. This technique plays a crucial role in training by preventing unknown outputs from destabilizing the training process [18]. Furthermore, we freeze the original layers to preserve their encoded knowledge, permitting only the copied layers to undergo training. These steps guarantee that the HuBERT model with BBE behaves identically to the HuBERT model without BBE during the initial stages. A high-level overview of ExHuBERT is depicted in Figure 2.

4. Experiments and Results

We split the experiments into two main parts: i) selection of the suitable audio Transformer for the BBE and fine-tuning of the chosen Transformer on EMOSET++ (cf. Section 4.1), and ii) conducting BBE on the fine-tuned Transformer and evaluating its performance on unseen emotion datasets (cf. Section 4.2).

4.1. Selection of the Suitable Audio Transformer for BBE

To choose the optimal architecture for BBE, we evaluate 6 state-of-the-art Transformers, including W2V2 XLS-R (300 million[1] and 1 billion[2]), Whisper (Medium[3] and Large v3[4]), and HuBERT (Large[5] and XLarge[6]) on all 26 emotion corpora of EMOSET [6]. We selected two variants of each architecture to compare their parameter impact, ensuring that variants of the same size had approximately equal parameter counts. Each Transformer is initialized with pre-trained weights obtained from `huggingface.co`. Additionally, we add a simple linear layer on top of each Transformer, with an output size of 6, corresponding to the mapped classes. For the evaluation metric, we use Unweighted Average Recall (UAR) due to its effectiveness in assessing the overall classification performance across all classes without bias towards any under- or oversampled class. We fine-tune all Transformer models in a round-robin fashion, sequentially passing each dataset forward and backward through the model with one batch in each step. For W2V2 and HuBERT variants, we use raw audio waveforms resampled to 16 kHz as inputs, while we feed Whisper with log Mel spectrograms (with either 80 or 128 bins) obtained from waveforms. We freeze the CNN encoder during the entire training process for W2V2 and HuBERT. This step is unnecessary for Whisper. We conclude the experimentation phase after 3k steps using AdamW optimisation with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-08$, and a learning rate of $1e-05$. The performance of these audio Transformers on the unseen test partition of eight commonly used emotion datasets is provided in Table 1[7].

Our results demonstrate that the HuBERT Large model outperforms others, achieving an average UAR of $62.7\,\%$ over all eight datasets, followed by W2V2 XLS-R 300 M with $57.1\,\%$ UAR. The worst-performing models are both variants of Whisper, each achieving $42.0\,\%$ and $41.1\,\%$ UAR, respectively. We consequently settle on HuBERT Large as the base model for fine-tuning on the extended EMOSET++ and transfer learning through BBE.

Subsequently, we test the impact of using both $83.7\,\%$ of EMOSET++ and the full EMOSET++ for fine-tuning the selected HuBERT model for speech emotion recognition, aiming to evaluate the impact of dataset size on model performance and generalization. Training on EMOSET++ leads to substantial performance gains for all databases, compared to the original EMOSET, raising the average UAR over the 8 databases from

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m
[2] https://huggingface.co/facebook/wav2vec2-xls-r-1b
[3] https://huggingface.co/openai/whisper-medium
[4] https://huggingface.co/openai/whisper-large-v3
[5] https://huggingface.co/facebook/hubert-large-ls960-ft
[6] https://huggingface.co/facebook/hubert-xlarge-ls960-ft
[7] Results on all datasets are provided here: https://huggingface.co/amiriparian/ExHuBERT/blob/main/supp-mat.pdf

Datasets	Performance in [% UAR]							
	W2V2 XLS-R 300M	W2V2 XLS-R 1B	Whisper Medium	Whisper Large v3	HuBERT Large	HuBERT XLarge	HuBERT Large 83.7 % **of EMOSET++**	HuBERT Large **EMOSET++**
Berlin Database of Emotional Speech (EMO-DB) [20]	**91.1**	86.0	55.0	61.2	90.0	82.4	88.1	**99.1**
Database of Elicited Mood in Speech (DEMOS) [21]	56.1	69.2	30.0	32.8	**67.7**	57.2	70.8	**90.7**
EmoFilm [22]	56.9	43.4	49.7	44.5	**58.6**	57.4	58.9	**60.5**
EmotiW-2014[23]	**40.0**	34.7	28.6	27.2	33.1	32.5	**40.2**	39.1
eNTERFACE [24]	66.4	76.1	39.3	38.6	**93.9**	63.2	92.9	**94.6**
IEMOCAP [25]	56.4	49.7	42.9	39.0	61.1	**65.8**	63.9	**67.8**
Multimodal EmotionLines Dataset (MELD) [26]	23.2	24.7	23.8	22.6	**30.0**	25.9	34.1	**38.5**
Mandarin Emotional Speech (MES) [27]	66.3	48.8	66.3	65.0	**67.5**	63.8	**70.0**	67.5
Average UAR over all datasets	57.1	54.1	42.0	41.4	62.7	56.0	64.9	69.7

Table 1: *Performance comparison of the applied Transformers on eight common speech emotion datasets. Our proposed fine-tuning of HuBERT Large on* EMOSET++ *demonstrates superior performance over all other Transformers. The best results without fine-tuning on* EMOSET++ *are bolded, and the best overall results (including fine-tuning on* EMOSET++*) are bolded and lightly shaded.*

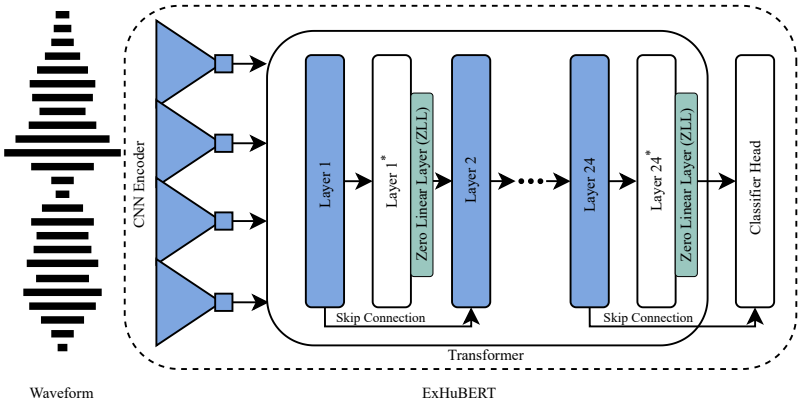

Figure 2: *Outline of the proposed ExHuBERT architecture, including skip connections and zero linear layers. The CNN encoder and the weights of the layers colored in blue are frozen. The classifier head and the weights of the layers marked with an asterisk (*) will undergo fine-tuning during the evaluation process.*

62.7 % to 64.9 % when only adding 5 training corpora, and to 69.7 % with the full set. All of the evaluated databases benefit from adding more corpora to EMOSET. However, EmoFilm and Mandarin Emotional Speech stop seeing gains after adding the first 5 new datasets.

4.2. ExHuBERT Comparison

We compare the transfer learning capabilities of our enhanced version of HuBERT (ExHuBERT Large EMOSET++) against other state-of-the-art Transformer models on a set of six unseen SER corpora, including Athens Emotional States Inventory (AESI) [30], Audio, Speech, and Vision Processing Lab Emotional Sound database (ASVP-ESD) [31], JL-Corpus [32], MLEnd[8], Synthesized Database of Basic Emotions (SyntAct) [33], and Variably Intense Vocalizations of Affect and Emotion Corpus (VIVAE) [34]. Specifically, we chose two variants of W2V2 – W2V2 XLS-R and the emotion fine-tuned W2V2 model by Wagner *et al.* [9] – and HuBERT Large LS960 to evaluate the impact of model architectures and the efficacy of EMOSET++ fine-tuning. Furthermore, we ablate the performance gains achieved through EMOSET++ fine-tuning from those due to block expansion by additionally expanding the LibriSpeech pre-trained HuBERT model (ExHuBERT Large LS960). Finally, we increase the number of trainable parameters of our ExHuBERT model by (1) unfreezing the original HuBERT layers (ExHuBERT Large Non-Frozen EMOSET++) and (2) tripling each original layer during block expansion (ExHuBERT XLarge EMOSET++). Table 2 shows the results and the number of trainable parameters for each of the evaluated models on six databases external to EMOSET++. For all but one of the six external databases, starting from a model that has been pre-trained on SER data, be it MSP-Podcast for W2V2 Emotion or EMOSET++ for HuBERT Large, leads to substantially improved performance, compared to the corresponding models trained on general speech data. The noteworthy outlier is found with SyntAct, which is a database of synthesized emotional speech. Here, performance has degraded from the respective base models of W2V2 and HuBERT large, indicating that fine-tuning on emotional speech data might have diminished generalization capabilities to non- and atypical SER tasks. However, BBE closes this performance gap, enabling efficient transfer from human-recorded to synthetic emotional speech. We hypothesize that without BBE, important SER knowledge acquired during pre-training gets overwritten early in the fine-tuning process due to the domain gap between synthetic and natural voices. Looking at the remaining databases, BBE leads to performance gains for HuBERT fine-tuned on EMOSET++, raising the average UAR from 71.1 % to 74.2 %. On the other hand, applying BBE to HuBERT pre-trained on LibriSpeech (ExHuBERT Large LS960) does not lead to con-

[8]https://mlenddatasets.github.io/spoken_numerals/

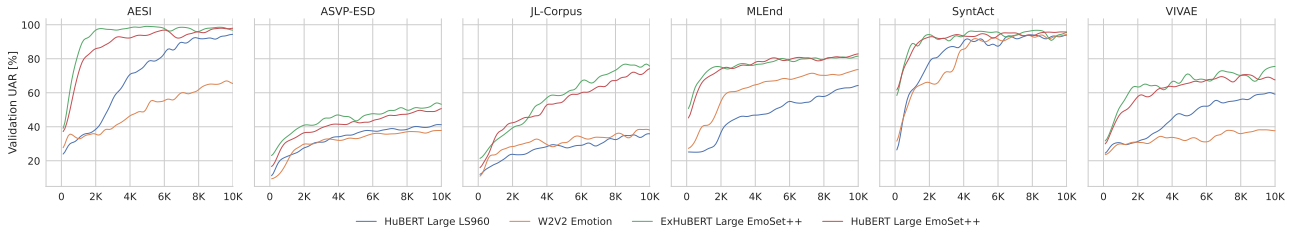

Figure 3: *Training curves on the six external SER corpora, displaying validation UAR.*

Dataset	Performance on unseen test in [% UAR]							
	W2V2 Emotion	W2V2 XLS-R	HuBERT Large LS960	HuBERT Large EMOSET++	ExHuBERT Large LS960	ExHuBERT Large EMOSET++	ExHuBERT Large Non-Frozen EMOSET++	ExHuBERT XLarge EMOSET++
AESI [30]	48.8	40.6	80.9	93.0	81.1	**94.7**	86.9	33.7
ASVP-ESD [31]	36.7	34.5	40.5	45.9	41.8	**51.4**	43.5	27.1
JL-Corpus [32]	41.4	36.8	47.9	66.7	45.6	**67.7**	47.7	41.7
MLEnd	73.6	65.2	74.6	**81.4**	77.7	80.0	77.9	60.1
SyntAct [33]	82.7	89.6	**93.7**	84.6	91.2	93.0	93.5	63.9
VIVAE [34]	45.0	45.6	42.1	54.5	46.4	**58.2**	49.3	25.1
Average UAR over all datasets	52.1	54.7	63.3	71.1	61.3	74.2	66.5	42.0
Number of trainable parameters	161.13 M	311.49 M	311.49 M	311.49 M	336.68 M	336.68 M	664.18 M	664.18 M

Table 2: *Performance comparison of the transfer learning capabilities of our proposed backbone extended HuBERT (ExHuBERT), and its variations, against state-of-the-art Transformers across six emotion datasets. None of these datasets were utilized in the fine-tuning process of* EMOSET++. *For each model, we also provide the average UAR over all datasets and the number of trainable parameters. The best results are bolded and lightly shaded.*

sistent improvements across the six SER corpora, with a slightly lower average UAR of 61.3 % compared to fine-tuning without BBE. Overall, BBE seems to have a positive effect on accuracy when the domain shift between source and target databases is rather small, e. g., from one SER corpus to another.

We further analyze how fast the different models converge during training on an unseen corpus in Figure 3, which shows the validation UARs over 10,000 training steps. Analogous to the test set results, EMOSET++ pre-training increases overall performance and further achieves faster convergence on all databases compared to models without SER pre-training and the MSP-Podcast fine-tuned W2V2. For AESI, BBE helps the ExHuBERT model to converge even earlier.

To conclude, we look at the results achieved with the larger versions of ExHuBERT, which double the amount of trainable parameters. Unfreezing the original HuBERT layers after BBE degrades performance even from simple fine-tuning of HuBERT on each of the target corpora, highlighting the importance of keeping the weights of the original model fixed for transfer learning. Expanding each block in ExHuBERT by adding a second copy of the respective layer suffers from substantial overfitting, achieving the worst overall UAR of all evaluated Transformer models.

In summary, the validation of our approach on both databases contained within EMOSET++ and external corpora shows that (1) multi-corpus pre-training on EMOSET++ leads to substantially increased SER performance on seen and unseen corpora, and (2) the addition BBE in ExHuBERT further helps with generalization to new datasets.

For running all of our machine learning experiments, we utilized one RTX-3090 GPU with 24 GB memory, and needed a total time of 313 hours: 121 h for stage 1 (pre-selection of Transformers), 87 h for EMOSET++ fine-tuning, 17 h for ExHuBERT EMOSET++ testing, and 88 h for the second stage.

5. Conclusions

We have proposed a novel twofold approach for SER by (i) collecting EMOSET++, a comprehensive multi-cultural and multilingual corpus comprising 37 emotion datasets, and (ii) introducing an enhanced version of HuBERT, denoted as ExHuBERT, achieved through fine-tuning on EMOSET++ and incorporating backbone extension. To find the suitable Transformer for BBE, we first selected six versions of state-of-the-art audio Transformers and analyzed their performance on EMOSET [6] and then fine-tuned the best-performing Transformer (which was HuBERT Large) on EMOSET++. In the subsequent phase, we applied BBE to the fine-tuned HuBERT Large model and compared its performance with other Transformers on six previously unseen emotion datasets. The experimental results underscore the effectiveness of our proposed approach, demonstrating its capacity to generalize across diverse datasets and establish new benchmarks for a variety of emotion recognition tasks. Lastly, we have uploaded ExHuBERT on `huggingface.co`[9]. Fine-tuning and deploying ExHuBERT may be computationally demanding, potentially **limiting** its use in resource-constrained environments.

For future work, we aim to include MSP-Podcast dataset [17] in EMOSET++ and enhance ExHuBERT for continuous recognition of arousal, valence, and dominance.

6. Acknowledgements

This work was supported by MDSI – Munich Data Science Institute as well as MCML – Munich Center of Machine Learning.

[9] https://huggingface.co/amiriparian/ExHuBERT

7. References

[1] J. D. Williamson, *Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person*, US Patent 4,093,821, Jun. 1978.

[2] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP*, IEEE, vol. 3, 1996, pp. 1970–1973.

[3] 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, Apr. 2018.

[4] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers in Neurorobotics*, vol. 15, 2021.

[5] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1912–1926, Jul. 2023.

[6] M. Gerczuk, S. Amiriparian, S. Ottl, and B. Schuller, "EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 13, 2022.

[7] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions," in *Proc. INTERSPEECH*, ISCA, Sep. 2019, pp. 3935–3939.

[8] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. INTERSPEECH*, ISCA, Sep. 2019, pp. 1691–1695.

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, Sep. 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[11] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, May 2022.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "SUPERB: Speech processing universal PERformance benchmark," in *Proc. INTERSPEECH*, ISCA, Aug. 2021, pp. 1194–1198.

[15] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Interspeech 2021*, pp. 3400–3404, Aug. 2021.

[16] Y. Wang, A. Boumadane, and A. Heba, *A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding*, Oct. 2022. arXiv: 2111.02735 [cs, eess].

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019.

[18] C. Wu, Y. Gan, Y. Ge, Z. Lu, J. Wang, Y. Feng, P. Luo, and Y. Shan, *LLaMA pro: Progressive LLaMA with block expansion*, Jan. 2024. arXiv: 2401.02415 [cs].

[19] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, ISCA, Sep. 2005, pp. 1517–1520.

[21] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, Jun. 2020.

[22] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Proc. INTERSPEECH*, ISCA, Sep. 2018, pp. 3638–3642.

[23] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol," in *Proc. ICMI*, New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 461–466, ISBN: 978-1-4503-2885-2.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in *Proc. ICDEW*, Apr. 2006.

[25] 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," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[26] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*, Jun. 2019. arXiv: 1810.02508 [cs].

[27] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[28] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus speech emotion recognition with hubert self-supervised representation," in *IberSPEECH*, ISCA, 2022, pp. 76–80.

[29] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[30] T. Chaspari, C. Soldatos, and P. Maragos, "The development of the Athens Emotional States Inventory (AESI): Collection, validation and automatic processing of emotionally loaded sentences," *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry*, vol. 16, no. 5, pp. 312–322, 2015.

[31] T. T. L. Dejoli, Q. He, and W. Xie, "Audio,Speech and Vision Processing Lab Emotional Sound database (ASVP-ESD)," May 2021.

[32] J. James, L. Tian, and C. Inez Watson, "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications," in *Proc. INTERSPEECH*, 2018, pp. 2768–2772.

[33] F. Burkhardt, F. Eyben, and B. Schuller, "SyntAct: A Synthesized Database of Basic Emotions," in *Proc. DCLRL*, J. Sälevä and C. Lignos, Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 1–9.

[34] N. Holz, P. Larrouy-Maestri, and D. Poeppel, "The variably intense vocalizations of affect and emotion (VIVAE) corpus prompts new perspective on nonspeech perception," *Emotion*, vol. 22, no. 1, pp. 213–225, 2022.