Interspeech 2024
1-5 September 2024, Kos, Greece

SER Evals: In-domain and Out-of-domain benchmarking for speech emotion recognition

Mohamed Osman[1], *Daniel Z. Kaplan*[2], *Tamer Nadeem*[1]

[1]Virginia Commonwealth University, United States
[2]realiz.ai, United States

osmanmw@vcu.edu, daniel.z.kaplan@realiz.ai, tnadeem@vcu.edu

Abstract

Speech emotion recognition (SER) has made significant strides with the advent of powerful self-supervised learning (SSL) models. However, the generalization of these models to diverse languages and emotional expressions remains a challenge. We propose a large-scale benchmark to evaluate the robustness and adaptability of state-of-the-art SER models in both in-domain and out-of-domain settings. Our benchmark includes a diverse set of multilingual datasets, focusing on less commonly used corpora to assess generalization to new data. We employ logit adjustment to account for varying class distributions and establish a single dataset cluster for systematic evaluation. Surprisingly, we find that the Whisper model, primarily designed for automatic speech recognition, outperforms dedicated SSL models in cross-lingual SER. Our results highlight the need for more robust and generalizable SER models, and our benchmark serves as a valuable resource to drive future research in this direction.

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

1. Introduction

Speech emotion recognition has garnered significant attention due to its potential to enable more natural and empathetic human-computer interaction. Recent advancements in self-supervised learning have led to powerful speech representation models like wav2vec2 [1], HuBERT [2], and WavLM [3], which have shown impressive performance on various speech processing tasks. However, the generalization of these models to diverse languages and emotional expressions remains a critical challenge [4].

Existing SER benchmarks often focus on a limited set of well-studied datasets, which may not accurately reflect real-world scenarios [5]. Moreover, the emphasis on in-domain evaluation fails to capture the crucial aspect of out-of-domain generalization, which is essential for deploying SER systems in practical applications. For our paper's purpose, we define in-domain as evaluating on the same data distribution seen in training, and out-of-domain as evaluating on a different data distribution. This can manifest as different speakers, tones, decision boundaries, etc. To address these limitations, we propose a large-scale benchmark that evaluates SER models on a diverse collection of multilingual datasets, emphasizing zero-shot performance.

Our benchmark focuses on less commonly used datasets to mitigate overfitting and encourage the development of more robust and adaptable models. We employ state-of-the-art speech representation models, including Whisper [6], an automatic speech recognition model, and CLAP [7, 8], a contrastive learning model, to analyze their performance in cross-lingual SER. Interestingly, our results show that Whisper consistently outperforms dedicated SSL models across most datasets, challenging the common belief that ASR models are suboptimal for SER due to their focus on phoneme recognition.

Table 1: *Multilingual datasets used in our benchmark. Values reflect the datasets after the class mapping.*

Dataset	Classes	Speakers	Language	Samples	Avg Duration (s)	OOD Eligible
URDUDataset[10]	4	6	Urdu	400	2.5	No
EmoDBDataset[11]	6	10	German	535	2.8	No
EMOVODataset[12]	7	10	Italian	588	3.1	Yes
eNTERFACEDataset[13]	6	42	English	1293	2.9	No
MESDDataset[14]	6	6	Spanish	1150	0.7	Yes
MASCDataset[15]	5	68	Mandarin	25636	1.9	No
DEMoSDataset[16]	7	68	Italian	9697	2.9	Yes
CASIADataset[17]	6	4	Mandarin	1200	1.9	No
AESDDDataset[18]	5	6	Greek	604	4.1	No
BAUMDataset[19]	8	31	Turkish	1398	4.6	Yes
EEKKDataset[20]	4	10	Estonian	1164	3.4	No
ThorstenDataset[21]	6	1	German	2399	4.4	No
RESDDataset[22]	7	200	Russian	1396	6.0	No
MELDDataset[23]	7	407	English	12924	3.2	Yes
MEADDataset[24]	7	60	English	31724	4.2	Yes
CaFEDataset[25]	7	12	French	936	4.4	Yes
ExpressoDataset[26]	8	4	English	11954	4.2	Yes
ShEMODataset[27]	6	87	Persian	3000	4.1	No
SUBESCODataset[28]	7	20	Bangla	7000	4.0	Yes

The main contributions of this work are as follows:

- We introduce a large-scale benchmark for evaluating the robustness and generalization of SER models across diverse languages and emotional expressions.
- We curate a collection of multilingual datasets and establish targetted subsets for systematic in-domain and out-of-domain evaluation.
- We employ logit adjustment[9] to account for varying class distributions and ensure fair comparisons across datasets.
- We conduct extensive experiments with state-of-the-art speech representation models and provide insights into their cross-lingual SER performance.
- We open source our entire code base, our full un-reduced results and training logs, as well as all implementation details at the following url: `https://github.com/spaghettiSystems/serval`.

2. Related Work

Self-supervised learning has revolutionized speech representation learning, enabling models to capture rich acoustic features without relying on labeled data. Models like wav2vec 2.0 [1], HuBERT [2], and WavLM [3] have achieved state-of-the-art performance on various speech processing tasks, including speech recognition, speaker identification, and emotion recognition [29].

Cross-lingual SER has gained attention as a means to develop models that can generalize across languages. Several studies have explored the use of SSL models for cross-lingual SER [30, 4]. However, these works often focus on a limited set of languages and datasets, making it difficult to assess the true

10.21437/Interspeech.2024-2440

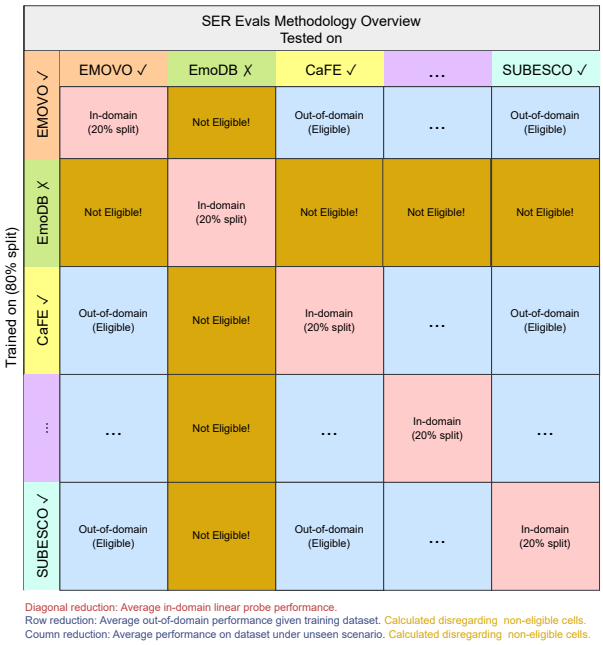

Figure 1: *Overview of our benchmark's methodology.*

generalization capabilities of the models.

Existing well-known SER benchmarks, such as IEMOCAP [31] and MSP-Podcast [32], have played a crucial role in advancing the field. However, these benchmarks often emphasize in-domain evaluation and may not adequately capture the challenges of real-world deployment [5]. Our work aims to address these limitations by introducing a large-scale benchmark that focuses on out-of-domain generalization and includes a diverse set of multilingual datasets.

Recent works such as EMO-SUPERB [33] and SERAB [34] have made notable contributions to the field of Speech Emotion Recognition (SER). However, these works have limitations in terms of the diversity of languages, datasets, and the emphasis on out-of-domain generalization.

Our benchmark significantly advances the state-of-the-art in SER evaluation by addressing these limitations. We curate an extensive collection of multilingual datasets, carefully selected to cover diverse linguistic and cultural contexts, ensuring a thorough evaluation of SER models in real-world scenarios. Moreover, our benchmark places a strong emphasis on out-of-domain generalization, a crucial aspect that has been largely overlooked in previous works. We evaluate SER models in both in-domain and out-of-domain settings, providing valuable insights into their ability to adapt to unseen data distributions. This focus on generalizability is essential for developing SER models that can be effectively deployed in real-world applications, where the variability in speech patterns, emotions, and recording conditions is vast.

3. Methodology

The primary objectives of this section are to detail the dataset selection and preprocessing steps, introduce the backbone models employed, describe the model architecture and training process, explain the logit adjustment technique, and outline the evaluation protocol. The methodology is designed to ensure a comprehensive and fair evaluation of state-of-the-art SER models across diverse languages and emotional expressions.

Table 2: *Backbone models used in our benchmark. All checkpoints are from Huggingface.*

Checkpoint name	Training Dataset Hours	# Params
facebook/w2v-bert-2.0 [35, 36]	4500k	580M
facebook/hubert-large-ll60k [2]	60k	315M
microsoft/wavlm-large [3]	94k	315M
laion/larger_clap_music_and_speech [7, 8]	$\gtrsim 10k$	193M
m-a-p/MERT-v1-330M [7]	160k	315M
openai/whisper-medium [6]	680k	307M
openai/whisper-large-v2 [6]	680k	636M
openai/whisper-large-v3 [6]	5000K	636M
openai/whisper-large [6]	680k	636M

3.1. Dataset Selection and Preprocessing

We curate a diverse collection of multilingual datasets for our benchmark, covering various languages and emotional expressions. Table 1 provides an overview of the datasets used in our evaluation. We focus on less commonly used datasets to mitigate overfitting and encourage the development of more robust models.

The datasets are preprocessed to ensure consistency and compatibility with our evaluation protocol. We set the maximum audio length to 30 seconds, and process the audios with appropriately for each backbone model we test (detailed in the next section). We rely on the Huggingface library for model preprocessing and inference implementations. Additionally, we remap the label space by mapping the original emotion labels to a unified eight-class space, facilitating cross-dataset comparisons. Due to complexity, detailing the exact remapping for each dataset is relegated to the open-source code.

The datasets used for out-of-domain evaluations are matched by having the same classes (excluding 'other') and their eligibility is indicated in the 'OOD Eligible' column of Table 1. These datasets were found to have the same exact classes after the class mapping, making them eligible for out-of-domain testing. When calculating out-of-domain metrics, samples with the 'other' label were discarded, and models were banned from predicting the 'other' class.

3.2. Backbone Models

We employ state-of-the-art speech representation models as backbones for our benchmark, as listed in Table 2. These models are selected based on their strong performance on various speech processing tasks and their ability to capture rich acoustic features.

In addition to the SSL models, we also evaluate MERT [37], a music recognition model, and CLAP [38, 8], a contrastive learning model. Including these models allows us to assess the effectiveness of different learning paradigms for cross-lingual SER. Lastly, we evaluate the Whisper[6] encoder which is trained under an encoder-decoder setup for ASR.

3.3. Model Architecture and Training

We employ a simple multilayer perceptron (MLP) architecture with approximately 500K parameters for emotion classification. The MLP consists of two hidden layers and is trained for 100 epochs. Due to the small parameter size and shallow depth, we do not expect substantial overfitting. We apply label smoothing with a factor of 0.1 to improve generalization.

Instead of the typical approach of averaging the features before classification, we execute the MLP on every feature frame and then take the mean of the predictions. We find that this approach preserves more information and leads to stronger and more consistent results.

3.4. Logit Adjustment

To account for the varying class distributions across datasets, we employ logit adjustment during evaluation. This technique adjusts the model's output logits based on the difference between the training and testing dataset distributions, mitigating the impact of class imbalance and enabling fair comparisons.

3.5. Evaluation Protocol

Figure 1 provides an overview of our benchmark's methodology. As we described in Subsection 3.1, we establish a subset of our datasets as OOD eligible, which have the same exact classes after the class mapping. Effectively, all datasets are accounted in in-domain tests. Only OOD-eligible datasets are accounted for our out of domain metrics.

For each model, we construct a performance matrix where the rows represent the training datasets and the columns represent the evaluation datasets. When the training and evaluation datasets are the same (diagonal elements), it indicates in-domain performance. Off-diagonal elements correspond to out-of-domain zero-shot performance.

We assess the quality of the backbone models based on three key metrics:

1. In-domain separability: We compute the mean of the diagonal elements to measure how well the features learned by a model can separate emotions within a dataset.
2. Out-of-domain performance given training dataset: We calculate the mean of each row, excluding the diagonal element, to evaluate the model's ability to generalize to unseen datasets when trained on a specific dataset.
3. Average performance on unseen datasets: We compute the mean of each column, excluding the diagonal element, to assess the average performance on a dataset when the model is not trained on it.

All metrics are reported in terms of macro-averaged F1 score to account for class imbalance.

4. Results and Discussion

The results of our benchmark provide valuable insights into the performance and generalization capabilities of state-of-the-art SER models across diverse languages and emotional expressions.

4.1. In-domain separability

The second and third column in Table 3 present the in-domain separability performance of various models, focusing on their ability to distinguish between different emotional states in speech. Performance is quantified by two metrics: the mean average performance across datasets (Mean) and the variability of performance across these datasets (Standard Deviation). From the table, Whisper-Large-v2 leads the evaluated models in in-domain SER performance, with the highest mean accuracy and low variability across datasets, closely followed by the original Whisper-Large. Other models like Whisper-Large-v3, Whisper-medium, WavLM-Large, and CLAP Music & Speech show competent but slightly more variable performances. Conversely, Hubert Large, MERT v1 330M, and w2v-bert-2.0 exhibit the lowest accuracies with higher fluctuations in their effectiveness across different datasets, indicating potential limitations in generalization capabilities for speech emotion contexts.

The outcome of this evaluation highlights a clear hierarchy among the models in terms of both accuracy and consistency in emotion recognition within the same domain. Whisper-Large variants stand out as the most effective, with their newer versions, particularly Whisper-Large-v2, slightly improving upon the original's already high benchmark. Lower-ranked models, though less consistent and accurate overall, may still offer valuable insights or perform well in specific niches or datasets. This analysis underscores the importance of choosing the right model for specific SER applications, balancing between performance and consistency across diverse emotional speech datasets.

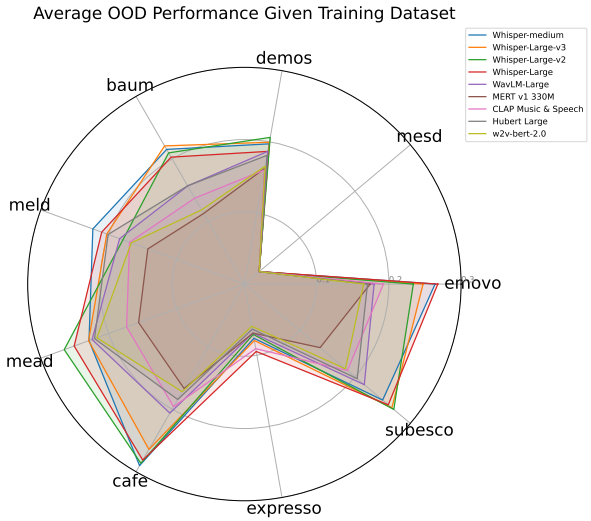

Figure 2: *Average out-of-domain performance given the training dataset.*

4.2. Out-of-domain performance given training dataset

Figure 2 shows the average out-of-domain performance for each model, obtained by row-wise reduction of the performance matrix. The Whisper models demonstrate the highest out-of-domain performance, indicating their superior generalization capabilities compared to the SSL models. However, there is high variability in OOD performance across training sets. Training on some datasets like BAUM leads to much better OOD generalization than others like MELD. This warrants further investigation into what properties of datasets lead to more generalizable models. The strong performance of Whisper challenges the common belief that ASR models are suboptimal for SER and highlights the potential of leveraging ASR models for emotion recognition tasks.

4.3. Average performance on unseen datasets

Figure 3 presents the average performance of the evaluated models on each dataset when the models are not trained on that dataset. The results highlight the varying levels of difficulty across datasets, with some datasets posing greater challenges for out-of-domain generalization. Notably, EMOVO, MELD, and MEAD are the most challenging for models not trained on them, suggesting they have unique characteristics that are harder to learn indirectly. On the other hand, models generalize best to URDU and AESDD, indicating these datasets share more common features with others. Interestingly, the Whisper model consistently achieves strong performance across most

Table 3: *Summary of key performance metrics for the evaluated models.*

Model	In-Domain (ID) Performance Average	In-Domain (ID) Performance Standard Deviation	Out-of-Domain (OOD) Performance Average	Out-of-Domain (OOD) Performance Standard Deviation	Weighted Average
Whisper-Large-v2	0.781942	0.194716	0.194250	0.089993	0.345741
Whisper-Large	0.781314	0.203542	0.197882	0.085657	0.344999
Whisper-Large-v3	0.776689	0.201399	0.192961	0.083822	0.342214
Whisper-medium	0.756563	0.200798	0.196831	0.087710	0.332443
WavLM-Large	0.765474	0.206174	0.161098	0.068182	0.326108
CLAP Music & Speech	0.743248	0.215404	0.148090	0.055976	0.309980
Hubert Large	0.733036	0.207492	0.156504	0.066509	0.307769
MERT v1 330M	0.707891	0.211471	0.127485	0.049841	0.287032
w2v-bert-2.0	0.668253	0.211685	0.141581	0.061820	0.268165

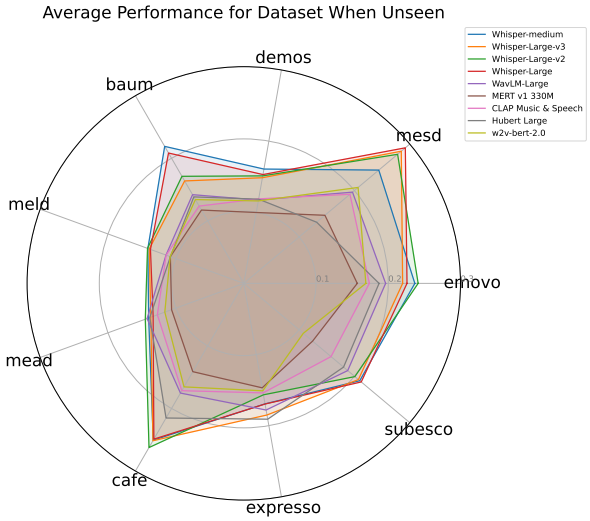

Figure 3: *Average performance on individual datasets when not trained on them.*

datasets, surpassing the SSL models in many cases.

4.4. General outcomes

Table 3 provides a summary of the key performance metrics for the evaluated models. The second and third columns show the average and standard deviations of the in-domain results, while the next two columns show the out-of-domain performance. The weighted average column is calculated as follows:

$$\text{Weighted Average} = \frac{\text{Average OOD} + \text{Average ID}}{2} - \lambda_{factor} \times \frac{\text{Std. Dev. OOD} + \text{Std. Dev. ID}}{2}$$

where λ_{factor} is the discounting factor, which we set to 1.0.

The Whisper models consistently achieve the highest scores across all metrics, further confirming their effectiveness in cross-lingual SER. However, the high standard deviations indicate that performance is quite variable depending on the specific train/test combination. This suggests that model robustness is still a challenge and there is room for improvement in developing models that perform consistently well across diverse datasets.

Our benchmark also demonstrates the effectiveness of logit adjustment in addressing the challenges posed by varying class distributions across datasets. By incorporating this technique, we ensure fair comparisons and mitigate the impact of class imbalance on model performance.

5. Conclusion

In this paper, we introduced a comprehensive benchmark for evaluating the robustness and generalization of speech emotion recognition models across diverse languages and emotional expressions. Our benchmark focuses on less commonly used datasets to mitigate overfitting and encourage the development of more robust models. Through extensive experiments with state-of-the-art speech representation models, we found that the Whisper model, primarily designed for automatic speech recognition, outperforms dedicated SSL models in cross-lingual SER. This finding challenges the common belief that ASR models are suboptimal for SER and highlights the potential of leveraging ASR models for emotion recognition tasks.

Our benchmark, along with the released code and evaluation protocol, serves as a valuable resource for the research community to assess and advance the state of cross-lingual SER. The insights gained from our work can guide future research efforts in developing more robust and generalizable SER models.

6. Future Works

Future directions include exploring advanced techniques for domain adaptation, few-shot learning, and meta-learning to further improve the generalization capabilities of SER models. Additionally, investigating the specific characteristics of datasets that contribute to better generalization can provide valuable insights for dataset design and selection.

We hope that our benchmark and findings will inspire researchers to push the boundaries of cross-lingual SER and develop models that can effectively handle the diversity of languages and emotional expressions encountered in real-world applications.

7. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] 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.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021.

[5] M. Osman, T. Nadeem, and G. Khoriba, "Towards generalizable ser: Soft labeling and data augmentation for modeling temporal emotion shifts in large-scale multilingual speech," *arXiv preprint arXiv:2311.08607*, 2023.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[7] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.

[8] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020.

[10] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *2018 International conference on frontiers of information technology (FIT)*. IEEE, 2018, pp. 88–93.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[12] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proc. LREC*, 2014.

[13] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd international conference on data engineering workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[14] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The mexican emotional speech database (mesd): elaboration and assessment based on machine learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1644–1647.

[15] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE Odyssey-the speaker and language recognition workshop*. IEEE, 2006, pp. 1–5.

[16] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, 2020.

[17] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for Mandarin text to speech," in *The Blizzard Challenge Workshop*, 2008.

[18] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 457–467, 2018.

[19] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[20] R. Altrov and H. Pajupuu, "Estonian emotional speech corpus: culture and age in selecting corpus testers," in *Human Language Technologies–The Baltic Perspective*. IOS Press, 2010, pp. 25–32.

[21] T. Müller and D. Kreutz, "Thorsten-voice dataset 2021.02," Sep. 2021, Please use it to make the world a better place for whole humankind. [Online]. Available: https://doi.org/10.5281/zenodo.5525342

[22] I. Lubenets, N. Davidchuk, and A. Amentes, "Aniemore." [Online]. Available: https://github.com/aniemore/Aniemore

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019.

[24] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. ECCV*, 2020.

[25] P. Gournay, O. Lahaie, and R. Lefebvre, "A Canadian French emotional speech dataset," in *Proc. ACM Multimedia*, 2018.

[26] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid *et al.*, "Expresso: A benchmark and analysis of discrete expressive speech resynthesis," *arXiv preprint arXiv:2308.05725*, 2023.

[27] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," in *Proc. LREC*, 2019.

[28] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," in *Proc. PloS One*, 2021.

[29] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[30] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, R. Schettini, and I. Shanin, "Semi-supervised cross-lingual speech emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121368, 2024.

[31] 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," in *Proc. LREC*, 2008.

[32] 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, pp. 471–483, 10 2019.

[33] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, "Emo-superb: An in-depth look at speech emotion recognition," *arXiv preprint arXiv:2402.13018*, 2024.

[34] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7697–7701.

[35] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[36] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[37] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[38] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.