Exploring the Impact of Model Architectures, Language-based Pre-finetuning, and Test Datasets on Speech Emotion Recognition

Anonymous submission to Interspeech 2024

Abstract

Speech Emotion Recognition (SER) is being increasingly applied in many societal contexts, often without adequate benchmarking. This study investigates the effects of model architectures, language-based pre-finetuning, and test datasets on the accuracy of SER systems, providing valuable insights for future SER studies and applications. We ran a statistical evaluation on two Italian emotional speech datasets (Emozionalmente and EMOVO), employing distinct self-supervised model architectures, both with and without Italian pre-finetuning. We found that model architectures and test datasets individually wield significant influence over the accuracy scores. Emozionalmente outperforms EMOVO with a highly significant difference, and Wav2vec 2.0 shows a similar level of significance in favor of HuBERT. Also, we found that model architectures, language-based pre-finetuning, and test datasets exhibit complex and interdependent interactions that collectively impact the accuracy of SER systems.

Index Terms: Speech emotion recognition, Italian, Self-supervised learning, Deep learning, Statistical analyses, Computational paralinguistics

1. Introduction

Speech Emotion Recognition (SER) involves identifying emotional attributes in speech independently of its semantic content [1]. This process involves the classification of emotions into distinct categories [2], with the widely recognized *Big Six* framework encompassing anger, fear, disgust, joy, sadness, and surprise, considered universally prevalent [3]. To automate SER, two approaches are prominent: one based on domain expertise, engineering handcrafted emotion-related features, and the other utilizing deep learning for data-driven representations [1]. Recently, self-supervised models, a type of deep learning approach, have shown promise in generating robust speech representations without labeled data, demonstrating impressive performance in SER and other audio tasks [4, 5]. Still, there remain some open questions. The literature presents conflicting recommendations regarding self-supervised model architectures, with some studies favoring Wav2vec 2.0 [6] and others advocating for HuBERT [7, 8]. Moreover, there is little understanding of the impact of language-based pre-finetuning [9] during model pre-training for downstream SER model performance [7]. How do they improve model accuracy or generalizability, and if so, is it worth the cost of additional training time?

Furthermore, in monolingual settings, prior research reveals a performance decline when utilizing SER models across distinct corpora within the same language [7, 10]. Notably, investigations have predominantly centered on English datasets, due to their abundance, leaving a gap in understanding the generalizability of findings across languages (including Italian) [7, 11, 8]. While English is one of the dominant languages in the world, we want research to benefit all, not just WEIRD [12] countries.

This study seeks to address these critical literature gaps by focusing on Italian. Our scope encompasses several key aspects, including comparative analyses of SER model performance on in-domain and cross-corpus test sets (Emozionalmente [13] and EMOVO [14], respectively), different architectural choices (i.e., Wav2vec 2.0 [15] and HuBERT [16]) and with and without Italian pre-finetuning. Additionally, we provide a framework for conducting similar investigations agnostic of language, architecture, and dataset[1].

2. Methodology

2.1. Datasets

Initially, we sought emotional speech datasets collected in natural settings. This approach was driven by the objective of ensuring that the emotional expressions contained within the data were as authentic and varied as possible, reflecting real-world scenarios. Despite our efforts, we could not find any, and this contrasts starkly with the availability of such datasets for other languages, exemplified by the MSP dataset [17] in the case of English.

We ended up using the EMOVO [14] and the Emozionalmente [13] datasets due to their shared design characteristics, such as both being in Italian, utilizing predetermined scripted sentences and encompassing the Big Six emotions [3] plus neutrality. Still, notable distinctions exist between the two datasets, including variations in recording settings and tools, size, speaker demographics, such as acting proficiency, and the way emotions are distributed across sentences and speakers (refer to Table 1 for detailed insights).

2.2. Models

We used contrastive (Wav2vec 2.0 [15]) and predictive (HuBERT [16]) self-supervised models to generate data-driven audio representations that could subsequently be employed for the emotion classification task. Specifically, we employed two pre-trained Wav2vec 2.0 "large" model checkpoints that use cross-lingual speech representations (XLSR) as the state-of-the-art for Wav2Vec architectures[18]. The first model[2] is a multi-language model pre-trained on 53 languages (including Italian) and was trained by the original creators of Wav2vec 2.0-XLSR [18], while the second one[3] is a mono-language model that

[1] https://github.com/anonymized_for_review

[2] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[3] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-italian

Table 1: *Overview of two Italian speech emotion datasets, EMOVO and Emozionalmente, highlighting key characteristics.*

Dataset	Recording setting	Recording tools	Truthfulness of emotion	Set of emotions	Linguistic content	Samples	Emotions and sentences balance	Speakers acting proficiency	Speakers mothertongue	Speakers sex distribution
EMOVO [14]	Laboratory	Professional microphone	Simulated	anger, disgust, fear, joy, sadness, surprise, neutrality	14 scripted sentences	588 audio recordings. Duration: M=3.12s, SD=1.36s	Every speaker acted out every sentence with every emotion	Professional	Italian	3M, 3F
Emozionalmente [13]	Variable (crowdsourced)	Variable (mostly commercial PCs', tablets' and smartphones' microphones)	Simulated	anger, disgust, fear, joy, sadness, surprise, neutrality	18 scripted sentences	6902 audio recordings. Duration: M=3.81s, SD= 0.99s	Every speaker acted out M=16, SD=22 sentences. Sentences were verbalized M=383, SD=15 times. Each emotion was acted out 986 times	Amateur	Italian	131M, 299F, 1 other

was obtained by pre-finetuning the aforementioned model using Italian data, aligning perfectly with our research focus on this language. We also employed two pre-trained HuBERT "large" model checkpoints. The former model[4] was pre-trained in English and was trained by the original creators [16], and the latter one[5] was obtained by pre-finetuning the former model on Italian language data, making it well-suited for our language-specific SER task. In terms of model architecture, the "large" version of HuBERT has the same configuration as the "large" version of Wav2vec 2.0, meaning they include 317 million parameters.

Notably, these two architectures were not pre-trained on the same initial datasets meaning that we will not be able to isolate the effects of pre-finetuning fully. However, transfer learning has become a dominant practice in the field and so the comparison of these two architectures and how pre-finetuning affects them will help researchers and developers choose between current pre-trained models and whether they should pre-finetune them.

2.3. Experimental setup

2.3.1. Data preparation

The loudness of the EMOVO and Emozionalmente datasets was normalized to -23 dB LUFS using `pyloudnorm` [19]. This step addresses the influence of loudness on emotion recognition, acknowledging its vulnerability to extraneous variables such as microphone distance [20]. By standardizing loudness, we ensure uniform audio levels across recordings, thereby reducing the influence of volume discrepancies on the auditory perception of emotional content [20].

To ensure robust training, accurate evaluation, and reliable generalization of our models, we implemented a rigorous data partitioning strategy for the Emozionalmente dataset. We carried out a two-stage 80-20 split, resulting in the creation of three distinct subsets: training, validation, and in-domain test, with approximate proportions of 64%, 16%, and 20%, respectively. To maintain a balanced representation of genders within each subset, we employed a stratified sampling technique. Furthermore, to guarantee complete independence among individuals in different sets, we implemented measures to ensure that any actor included in one subset was strictly excluded from the others. The EMOVO dataset was used as the cross-corpus test set to test model generalizability in very different conditions from that used during training.

[4]https://huggingface.co/facebook/hubert-large-ll60k

[5]https://huggingface.co/jonatasgrosman/exp_w2v2t_it_hubert_s722

2.3.2. Models training procedure

We finetuned the Wav2vec 2.0 and HuBERT checkpoints from Section 2.2 on the training subset of Emozionalmente for SER. This fine-tuning procedure included the mounting of a simple emotion classifier on top of the pre-trained models.

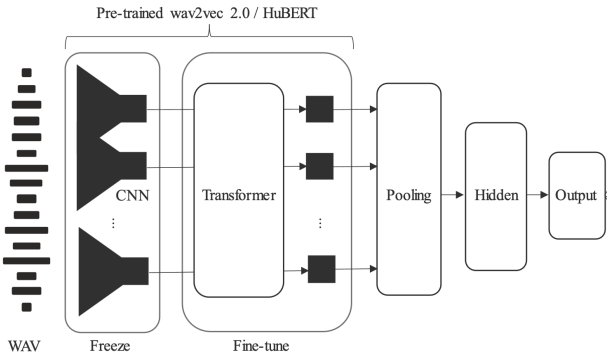

Figure 1: *Pretrained model architectures with a classification head showing the full model that was fine-tuned. Adapted from [7] Fig 1.*

Since Wav2vec 2.0 and HuBERT cannot inherently form an utterance-level audio representation, we decided to finetune them on a speech emotion classification task at an utterance level. We experimented with different architectures to put on the top of the model. Ultimately, we selected a configuration involving global average pooling across the time dimension to accommodate audio recordings of varying lengths. This was followed by the ReLU activation and a singular linear layer for emotion category prediction (as seen in Figure 1), a strategy aligned with the authors in [21]. Fine-tuning was performed using an A100 GPU. Our choice of a batch size of 16 was guided by computational resource constraints and memory availability. The training process encompassed 50 epochs, with early stopping employed to prevent overfitting. Early stopping was triggered by monitoring the validation accuracy, with a patience of 25 evaluations and a delta of 0.01.

2.3.3. Statistical evaluation

The evaluation of the trained models was performed on the Emozionalmente test subset and the entire EMOVO dataset. We focused on assessing the unweighted accuracy of the models, considering the balanced distribution of emotions in the

dataset. To ensure the statistical robustness of our findings, we ran 110 seed-based experiments to gauge the average performance across multiple runs. The determination of the number of seeds was guided by an a priori power analysis through the G*Power toolkit [22]. In Section 3, we present the results of a *3-way ANOVA*, followed by subsequent post-hoc *t-tests*, to test the following null-hypotheses:

1. There is no significant difference in the accuracy scores due to different architectures.
2. There is no significant difference in the accuracy scores due to the Italian pre-finetuning.
3. There is no significant difference in the accuracy scores due to different test sets.
4. There is no interaction effect on the accuracy scores between the architectures, the Italian pre-finetuning, and the test sets.

3. Results

We evaluated the accuracy of two different model architectures, namely Wav2vec 2.0 and HuBERT, with and without Italian pre-finetuning, on two test sets, i.e., Emozionalmente and EMOVO. Table 2 shows the accuracy metrics across 110 seeds for each of the tested conditions.

A *3-way between groups ANOVA* was used to examine the main effects and interactions of the Italian pre-finetuning, the architectures, and the test sets as they relate to the model accuracy. The results revealed that the three-way interaction was not statistically significant ($F(1, 872) = 3.5519$, $p = 0.0598$). Therefore, *the combined effect of these factors on accuracy is not significantly different from what one would expect based on the individual and pairwise effects*.

A significant 2-way interaction effect emerged between Italian pre-finetuning and architectures, as denoted by $F(1, 872) = 6.7778$, $p < 0.01$. Additionally, a significant interaction was identified between Italian pre-finetuning and test sets, $F(1, 872) = 6.9061$, $p < 0.01$. Furthermore, a highly significant interaction was evident between architectures and test sets, $F(1, 872) = 86.8540$, $p < 0.001$. These findings collectively indicate that *there are complex and interdependent relationships among Italian pre-finetuning, different architectural choices, and the selection of test sets, suggesting that these factors taken 2 by 2 influence the accuracy scores of the study*.

Simple main effects analysis showed that the Italian pre-finetuning had a statistically significant effect on accuracy ($F(1, 872) = 7.5911$, $p < 0.01$. Also, a substantial main effect was observed for different architectures, as evidenced by $F(1, 872) = 431.2770$, $p < 0.001$. Moreover, the test sets employed in the study exerted a statistically significant influence on accuracy scores, as reflected by the $F(1, 872) = 15002.5765$, $p < 0.001$. These findings suggest that *the Italian pre-finetuning, architectural choices, and the selection of test sets, singularly exert notable impacts on the accuracy outcomes within the context of the study*.

In a post-hoc analysis, we conducted *two-tailed t-tests* to investigate three key questions. Firstly, we examined whether there was a statistically significant difference in accuracy between the Emozionalmente and EMOVO datasets. Secondly, we assessed whether the accuracy of the models exhibited significant disparities when they were pre-finetuned in Italian. Lastly, we investigated whether there were significant variations in model accuracy when employing different architectural choices, specifically HuBERT and Wav2vec 2.0. Detailed results can be found in Table 3. Notably, *Emozionalmente outperforms EMOVO with a highly significant difference, and Wav2vec 2.0 shows a similar level of significance in favor of HuBERT. However, the comparison between Italian pre-finetuning and no language-based pre-finetuning indicates no statistically significant difference*.

4. Discussion

4.1. The choice of a self-supervised model architecture significantly impacts the accuracy of SER systems.

We observed that the Wav2Vec 2.0 architecture outperformed HuBERT for SER tasks. This finding aligns with the results reported in a similar study [6] conducted using an English emotional speech corpus, but it contradicts the conclusions drawn in the literature reviews in [7, 8] on the same dataset. It is important to notice that none of the referenced studies conducted a statistical analysis to measure the effect size of the results. Further analysis is needed to assess the generalizability of our results across different languages.

One potential explanation for the performance difference between Wav2Vec 2.0 and HuBERT could be related to the type of pretext task used during their training process. Predictive approaches, such as HuBERT, self-generate labels through statistical analyses of the speech signal and design prediction-based pretext tasks based on these self-generated labels. On the other hand, contrastive learning models, such as Wav2Vec 2.0, aim to distinguish target samples from distractor samples by minimizing latent space distances between positive samples and maximizing distances to negative samples. This second approach encourages audio representations with inter-class separability and intra-class compactness, which can benefit downstream classifier learning [23].

Additionally, one may speculate that the improved performance of the Wav2Vec2-XLSR model may be attributed to its ability to create cross-linguistic speech representations. This is likely due to the model being pre-trained on a multilingual dataset, which allows it to capture linguistic variations and nuances that could be beneficial for SER tasks. Further research is needed to confirm this idea.

4.2. Language-based pre-finetuning has minimal impact on SER model accuracy.

While the *ANOVA* test indicates that language-based pre-finetuning does have a statistically significant effect on model accuracy, this effect is relatively small. A further investigation (i.e., a *t-test* between the accuracies of all models evaluated with and without language-based pre-finetuning) reveals no statistically significant effects of pre-finetuning on model accuracy. These results seem to suggest that the positive influence of language familiarity on emotion recognition from speech, as reported in the behavioral literature [24], may not necessarily extend to machine learning models.

It is worth noting that our investigation is confined to the specific context of pre-finetuning on Italian. It is possible that languages with fewer emotionally labeled linguistic resources may yield more substantial benefits from this process. Moreover, our analysis has focused exclusively on accuracy as the performance metric, while other factors, such as training time and the volume of training data, should also be taken into account in future research.

Furthermore, it is important to acknowledge that our comparison involves two distinct models, Wav2vec 2.0 and HuBERT, both pre-trained on a substantial amount of data (56k and

Table 2: *Accuracy metrics for two different models across two emotion recognition datasets. The models were evaluated with and without language-based pre-finetuning. Accuracies, in decimal format, were computed over 110 random seeds.*

Model		**Test set**							
Architecture	**pre-finetuning**	**Emozionalmente**				**EMOVO**			
		Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Wav2vec 2.0	Italian pre-finetuning	0.7825	0.0166	0.7233	0.8328	0.5855	0.0341	0.4745	0.6854
	No language-based pre-finetuning	0.7810	0.0195	0.7313	0.8238	0.5675	0.0338	0.4796	0.6429
HuBERT	Italian pre-finetuning	0.7604	0.0170	0.7108	0.7999	0.5227	0.0347	0.4320	0.5935
	No language-based pre-finetuning	0.7615	0.0181	0.7068	0.7974	0.5211	0.0309	0.4439	0.5884

Table 3: *Comparison between Condition 1 and Condition 2 in terms of accuracy measurements, represented by Mean and Std across 110 seeds. The statistical significance of the differences is denoted by asterisks: *** means p<0.001.*

Condition 1	**Condition 2**	**Mean 1**	**Std 1**	**Mean 2**	**Std 2**	**N**	**t**	**p-value**	**Significance**
Emozionalmente	EMOVO	0.7714	0.0206	0.5492	0.0436	880	96.4847	0.0000	***
Wav2vec 2.0	HuBERT	0.6791	0.1064	0.6415	0.1224	880	-4.8724	0.0000	***
Italian pre-finetuning	No language-based pre-finetuning	0.6628	0.1145	0.6578	0.1179	880	-0.6380	0.5236	

60k hours, respectively). However, their pre-training data differ in terms of linguistic diversity, with one being multilingual and the other monolingual. To gain a comprehensive understanding of the effects of pre-finetuning, it is imperative to explore the broader implications of the specific languages involved in both pre-training and pre-finetuning processes.

4.3. Self-supervised SER models exhibit significantly different performance on in-domain and cross-corpus test sets.

In particular, in our experiments, Emozionalmente significantly outperforms EMOVO. While it may not come as a surprise that our results exhibit a drop in accuracy on the cross-corpus EMOVO dataset, considering our utilization of Emozionalmente as both the training and in-domain test set, this outcome raises pertinent questions concerning the role of different test sets in SER. The observed decline in accuracy aligns with prior SER experiments that employed cross-corpus testing data [7, 10], underscoring the challenge of adapting these systems to different domains and for real-world applications.

Though our findings strongly support the use of Emozionalmente as an Italian SER dataset, this current study does not fully rule out the use of EMOVO as a training set. Preliminary work [25] showed that the data in EMOVO is insufficient for robust speech emotion recognition performance with the proposed method (25% accuracy on Emozionalmente as the test set); however, we leave it to our future work to verify this under strict statistical analyses.

Transformer-based models have emerged as a promising direction, seemingly outperforming non-transformer baselines [7]. Nonetheless, as we showed, they are not without limitations. It is imperative that the research community advances in this direction to ensure that the models developed can facilitate applications beyond laboratory environments. For monolingual models, the inclusion of cross-corpus test sets remains crucial for evaluating their capacity to generalize. Nevertheless, this endeavor poses challenges, in particular given the scarcity of SER data, especially in languages beyond English. One promising avenue for future research lies in multi-lingual models [26], as they inherently necessitate generalized capabilities to excel across various languages. However, such models still require refinement.

5. Conclusion

In our study, Wav2Vec 2.0 demonstrated superior performance over HuBERT in Italian SER. Language-based pre-finetuning did not impact model accuracy significantly. We observed limitations in SER model generalization to a cross-corpus test set, highlighting the ongoing challenges in representing voice in a universally effective manner within the SER domain. We leave the validation of our findings across different languages to future work. Finally, we advocate for the need of statistical analyses to validate and contextualize research findings, with the hope of encouraging this practice among future researchers in the SER field, and the broader field of machine learning.

6. References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, p. 90–99, Apr. 2018. [Online]. Available: https://doi.org/10.1145/3129340

[2] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proc. of the national academy of sciences*, vol. 114, pp. 7900–7909, 2017.

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, pp. 169–200, 1992.

[4] J. Turian *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[5] S. W. Yang, P. H. Chi, Y. S. Chuang, C. I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi *et al.*, "Superb: Speech processing universal performance benchmark," in *INTERSPEECH 2021*. ISCA, 2021, pp. 3161–3165.

[6] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022*. IEEE, 2022, pp. 6922–6926.

[7] 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 Trans. on Pattern Analysis & Machine Intelligence*, pp. 1–13, 2023.

[8] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and evaluating speech emotion recognition systems: A reality check case study with iemocap," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt lan-

guage models to domains and tasks," in *Proc. of the 58th Annual Meeting of the ACL*, 2020, pp. 8342–8360.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. on Affective Computing*, 2022.

[11] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools and Applications*, pp. 1–68, 2021.

[12] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" *Behavioral and brain sciences*, vol. 33, no. 2-3, pp. 61–83, 2010.

[13] F. Catania, "Speech emotion recognition in italian using wav2vec 2.0 and the novel crowdsourced emotional speech corpus emozionalmente," *TechRxiv*, 2023.

[14] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *LREC 2014*. ELRA, 2014, pp. 3501–3504.

[15] 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.

[16] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[17] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech 2020*, 2020.

[18] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv:2006.13979*, 2020.

[19] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[20] G. Zhang, S. Qiu, Y. Qin, and T. Lee, "Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis," in *ISCSLP 2021*. IEEE, 2021, pp. 1–5.

[21] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[22] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," *Behav. Res. Methods*, vol. 41, pp. 1149–1160, Nov. 2009.

[23] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. on knowledge and data engineering*, vol. 35, pp. 857–876, 2021.

[24] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, p. 203, 2002.

[25] F. Catania, "Designing and engineering emotion-aware conversational agents to support persons with neuro-developmental disorders," 2022.

[26] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus training strategy for speech emotion recognition using self-supervised representations," *Applied Sciences*, vol. 13, p. 9062, 2023.