Bridging Emotions Across Languages: Low Rank Adaptation for Multilingual Speech Emotion Recognition

Lucas Goncalves[1], *Donita Robinson*[2], *Elizabeth Richerson*[2], *Carlos Busso*[1]

[1]Department of Electrical and Computer Engineering, The University of Texas at Dallas, USA
[2]Laboratory for Analytic Sciences, North Carolina State University, USA

{goncalves,busso}@utdallas.edu,{drobins7,ericher}@ncsu.edu

Abstract

The field of *speech emotion recognition* (SER) is constantly evolving with the surge in voice data and linguistic diversity. This growth highlights the need for SER systems capable of overcoming language barriers in both linguistic structure and cultural expression of emotions. We envision a SER framework that captures general trends in the expression of emotions, while also modeling language-specific information. Our study investigates *low rank adaptation* (LoRA) for creating multilingual SER models, applying LoRA in a multilingual context to efficiently adapt pre-trained models to new languages with minimal changes. This enhances cross-lingual adaptability and efficiency of SER systems, refining models to recognize emotions across languages without extensive retraining. In this study, we focus on exploring this method to bridge the gap between English and Taiwanese Mandarin in naturalistic settings, demonstrating strong performance in both languages.

Index Terms: speech emotion recognition, low rank adaptation, cross-lingual adaptability, affective computing

1. Introduction

With global communication increasingly crossing linguistic boundaries [1], it is important to design *speech emotion recognition* (SER) systems capable of functioning across multiple languages. This need is further intensified by the considerable differences in linguistic structures and cultural nuances of the expression of emotions [2, 3], especially among languages from different families. Feraru et al. [4] investigated SER's performance across various languages, distinguishing between within-language family and cross-language family scenarios. The findings indicated the necessity of selecting a base language closely related to the target languages for optimal SER model performance. This observation underscores the complexities of developing systems for cross-lingual applications, particularly for languages that are very distinct, such as English and Taiwanese Mandarin. These languages vary not only in phonetic and syntactic structures but also in their use of prosody to convey lexical and emotional information, posing significant challenges for conventional SER models [5].

Traditional SER systems, primarily designed for monolingual use, often struggle to recognize emotions in languages other than their training language, as evidenced by previous studies [4, 6, 7]. These cross-lingual challenges limit the effectiveness of SER systems in multilingual contexts and highlight a significant gap in the field's capability to serve a globally diverse audience. Recognizing this gap, our study explores the use of *Low Rank Adaptation* (LoRA) [8] to create versatile multilingual SER models. LoRA delivers a *parameter-efficient fine-tuning* (PEFT) approach to adapt models to perform better on tasks diverging from the original main task a model has been trained for. PEFT approaches like embedding prompt learning [9], adapter tuning [10], and LoRA [8] are techniques that only require a small number of parameters to be updated for adaptation, taking advantage of already pre-trained networks. Approaches such as LoRA have been widely explored in natural language processing [11, 10]; however, these approaches have not been widely explored for SER [12].

This study explores the applicability of LoRA to speech emotion recognition specifically focused on modeling emotional attribute prediction in a multilingual setting. A key difference in our study is the use of LoRA to introduce information about a second language into a monolingual model by tuning only a fraction of the parameters needed for a full fine-tuning process. The goal is to build a model that is language agnostic and retains strong performance across languages without the need to specify the target language when processing inputs by the emotion recognition model. As PEFT has been shown to perform well in other domains like vision [13], speech [14], and categorical SER [12], we expect to capture language-specific information using LoRA, complementing the information provided by the core language-agnostic representation. Furthermore, we also analyze how effectively we can achieve robust performance in more than one language with a single model that has been trained with two very dissimilar languages.

We validate our experiments on two naturalistic datasets: the MSP-Podcast corpus [15] for English and the BIIC-Podcast corpus [16] for Taiwanese Mandarin. We assess the effectiveness of our proposed method by implementing our ideas with two large variants of pre-trained *self-supervised learning* (SSL) frameworks, WavLM [17] and HuBERT [18]. Our findings indicate that our proposed method for multilingual SER with LoRA enhances cross-lingual performance while maintaining robust results in the original language used to initially fine-tune the SSL models. Our experimental results show that this approach is capable of maintaining a balanced performance across emotional dimensions in a cross-lingual context. Additionally, our study examines the impact of data volume on the cross-lingual efficacy of these models when applying LoRA. The approach can achieve comparable results with a significantly reduced number of recordings – 60% less for the BIIC-Podcast corpus (Taiwanese Mandarin) and 76% less for the MSP-Podcast corpus (American English) – demonstrating the potential of LoRA for efficient language adaptation in SER.

2. Background

2.1. Multilingual Emotion Recognition

Early studies conducted research in multilingual SER, examining its efficacy across a range of languages [19, 4, 20]. Feraru et al. [4] differentiated between language families. The study showed that maintaining consistent performance across diverse

languages presents additional challenges, finding that the use of a base model trained on a language with linguistic similarities to the target languages significantly enhances SER model performance. Lee et al. [21] explored the impact of employing regularization and normalization techniques in SER for heterogeneous languages. Their findings indicated that using a multitask learning framework, which simultaneously predicts gender, emotion, and language as auxiliary tasks, improves SER models' performance in multilingual contexts. Zehra et al. [22] showed that using an ensemble approach with majority voting is a viable method for multilingual cross-corpus SER.

Sharma [23] investigated the development of a multilingual SER system utilizing the pre-trained Wav2vec 2.0 model, fine-tuning it across multiple datasets to recognize categorical emotions effectively. The results showed that the model was able to outperform state-of-the-art methods for languages contained in the pre-training corpora. More recently, Upadhyay et al. [24] explored implementing domain adaptation in cross-lingual scenarios using phonetic constraints. The study found that identifying emotion-specific phonetic commonality across languages and using common vowels as phonetical anchors to perform unsupervised cross-lingual SER leads to strong performance in cross-lingual settings.

2.2. Fine-Tuning for Speech Emotion Recognition

The SER field has considerably evolved in recent years with the rise and availability of SSL models. A common strategy is to fine-tune a pre-trained model for SER, which has led to clear performance improvements [25, 26]. Initial studies, such as those by Pepino et al. [27], explored the efficacy of Wav2vec 2.0 embeddings [28] for SER. Morais et al. [29] extended this work by fine-tuning both Wav2vec 2.0 and HuBERT [18]. In these studies, a common trend was the superior performance over traditional feature-based methods achieved by fine-tuning SSL models. Wagner et al. [25] offered a thorough analysis of fine-tuning strategies for SER using Wav2vec 2.0 and HuBERT.

Feng and Narayanan [12] moved away from a fine-tuning strategy for SER. Instead, they presented an extensive analysis using PEFT on pre-trained speech models for categorical SER. In contrast, we present a novel use of PEFT using LoRA for the prediction of emotional attributes in the context of multilingual SER, effectively capturing language-specific information.

3. Methodology

We envision a global language-agnostic SER framework that is complemented by adjusting a smaller set of parameters that captures language-specific information. We implement our strategy with *Low Rank Adaptation* (LoRA). The overall structure of our approach is depicted in Figure 1 and comprises three primary modules. The *Encoder* ("Enc.") module incorporates components from the pre-trained SSL models, such as WavLM [17] or HuBERT [18]. Within our framework, this module undergoes fine-tuning with data from a source language, which we refer to as *language 1*. The LoRA module involves blocks that utilize lower-rank adaptation matrices for reparametrization; here, only A and B are trained with the target language, which we refer to as *language 2*. The combination of parameters from the "Enc." and "LoRA" blocks is performed as shown in Eq. 1,

$$h = (W_0)_{language1}\, x + (BA)_{language2}\, x \qquad (1)$$

where x is the input, W_0 represents the weights of the larger pre-trained model (blue box in Fig. 1 - Enc.) with matrix $W_0 \in \mathbb{R}^{d\times k}$ and BA is the LoRA weights (orange box in Fig. 1 - LoRA) with matrices $B \in \mathbb{R}^{d\times r}$, and $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$.

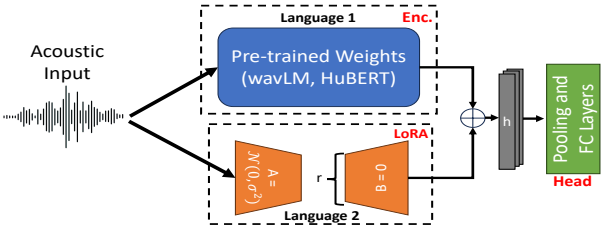


Figure 1: *Approach overview. Encoder ("Enc.") with pre-trained weights from SSL models (blue) is fine-tuned with language 1. LoRA weights (orange) is subsequently trained with language 2. Pooling layers and FC layers (green) at the head of the model receive information from both languages at once.*

The third block of our approach is the "Head," at the back-end stage. We employ pooling layers followed by *fully-connected* (FC) layers for the final prediction. These layers are the only layers that are trained and optimized with either one or both languages. The pooling layers leverage attentive statistics pooling [30], which applies an attention mechanism to assign varying weights to different frames, thus calculating both weighted means and weighted standard deviations for the frames outputted from the combination of the "Enc." and "LoRA" outputs. The resulting output from the pooling layer is then passed on to a simple set of FC layers for the emotional attribute predictions.

4. Experimental Settings

4.1. Resources

In our experiments, we leverage two public corpora derived from naturalistic recordings: the MSP-Podcast [15] and the BIIC-Podcast [16] corpora.

The MSP-Podcast corpus [15] serves as our primary source for English-speaking data. This study utilizes version 1.11 of the corpus, which contains 151,654 speaking turns from a variety of audio recordings, all under Creative Commons licenses. The training set is composed of 84,030 speaking turns. Additionally, the corpus contains a development set with 19,815 segments. For this study, we utilize test set 1, which contains 30,647 segments. At least five annotators have assessed each speaking turn, providing annotations on emotional attributes, as well as primary and secondary emotional categories. This study focuses on the emotional attributes of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong). The corpus annotations have a Krippendorff's alpha coefficient (α) inter-annotator agreement of 0.439 for arousal, 0.384 for dominance, and 0.505 for valence. We employ the BIIC-Podcast corpus [16] for Taiwanese Mandarin-speaking data, utilizing version 1.01 of this corpus. It includes 48,815 utterances in the training set, with additional sets for development (10,845 utterances) and testing (10,340 utterances). Similar to the MSP-Podcast corpus, each utterance in the BIIC-Podcast corpus is annotated by at least five annotators. These annotators provide evaluations for primary and secondary emotions, and emotional dimensions for arousal, valence, and dominance, which are the focus of our study. The corpus annotations have an inter-annotator agreement of 0.418 for arousal, 0.432 for dominance, and 0.461 for valence.

Table 1: *In-domain results comparisons. CCC scores are reported on the language the models were originally trained on.*

	MSP-Podcast			BIIC-Podcast		
HuBERT						
	Aro.	Dom.	Val.	Aro.	Dom.	Val.
Off-the-shelf	0.632	0.549	0.384	0.523	0.121	0.280
FineTuned	0.683	0.619	0.645	0.616	0.093	**0.382**
LoRA	**0.684**	0.614	0.641	0.613	0.079	0.376
FT(dual)	0.667	0.594	0.595	0.598	**0.188**	0.305
FT+LoRA	0.682	**0.620**	**0.646**	**0.617**	0.156	0.352
WavLM						
	MSP-Podcast			**BIIC-Podcast**		
Off-the-shelf	0.639	0.564	0.461	0.575	0.186	0.277
FineTuned	0.687	**0.619**	0.648	0.585	0.148	0.347
LoRA	0.685	0.616	**0.651**	0.601	0.104	0.329
FT(dual)	**0.689**	0.612	0.642	0.621	0.132	**0.383**
FT+LoRA	0.686	0.603	0.649	**0.622**	**0.184**	0.307

4.2. Implementation Details

We use two pre-trained SSL models to conduct our experiments. Specifically, we are using the large versions of WavLM [17] and HuBERT [18]. These models contain 24 transformer layers and are comprised of ∼310M parameters. We used these models for our study since they are the highest-ranked SSL models with this configuration for emotion recognition on the SUPERB Benchmark [31]. We utilized the pre-trained off-the-shelf models from Hugging Face [32] "facebook/hubert-large-ll60k" for HuBERT and "microsoft/wavlm-large" for WavLM. As evidenced in [25], fine-tuning SER models from pre-trained SSL models can lead to a significant boost in performance. The first step is to fine-tune these models using *language 1* (blue and green blocks in Figure 1). We fine-tune these models for 30 epochs, with a learning rate set to 1e-5, batch size of 32, and Adam as our optimizer. We use a multi-task setup, where we jointly predict arousal, valence, and dominance, all at once. The loss function for the emotional attributes in the regression models relies on the *concordance correlation coefficient* (CCC), which measures the agreement between the true and predicted emotional attribute scores. Equation 2 illustrates the CCC measurements,

$$\mathcal{L}_{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

where μ_x and μ_y represent the means of the actual and predicted scores, respectively, and σ_x and σ_y denote the standard deviations of these scores. The term ρ corresponds to the Pearson correlation coefficient between the true and predicted scores. The training objective is to achieve a high correlation between predicted and actual scores while minimizing prediction errors.

The second part of our model training involves incorporating LoRA weights into the framework for training. At this stage, we freeze the weights learned previously from *language 1* by the pre-trained SSL model layers (blue block in Fig. 1). Then, we train the model with only *language 2*, where the trainable sections of the model are set to be the orange and green blocks, as depicted in Figure 1. Our LoRA setup is configured with a rank of 32, the alpha parameter is set to 64, and the dropout rate is set to $p = 0.05$. We specify the following modules of the transformer layers from the SSL model to be targeted by the LoRA fine-tuning step: *Query* (Q), *Key* (K), and *Value* (V) matrices, output projection, intermediate dense layers, and output dense layers. The use of LoRA allows to reparametrize the model by only training a small percentage of the total model weights. Our complete model with the pre-trained SSL model, LoRA weights, pooling layer, and FC layers contains approximately 329M parameters. When applying LoRA, we only update ∼14M parameters, which is just 4.3% of the model's parameters. Similar to the training setup with *language 1*, we train the LoRA weights with *language 2* for 30 epochs, with a learning rate set to 1e-5, batch size of 32, and Adam as our optimizer. Everything was implemented in PyTorch and trained using an NVIDIA A100 GPU with 40GB.

Table 2: *Cross-Lingual CCC results comparisons. We conduct a one-tailed Welch's t-test between the approaches to assert significance at p-value ≤ 0.05. The symbol ∗ indicates statistical significance over the highest performing model in each attribute for in contrast to the other models. "M" represents the MSP-Podcast corpus and "B" represents the BIIC-Podcast corpus. In the conditions "M/B" or "B/M," the order in which they are listed represents the order in which we used the datasets during the training process. The symbol "—-" indicates that the specific module in the respective column was not used.*

					MSP-Podcast Test		
		Head	Enc.	LoRA	Aro.	Dom.	Val.
HuBERT	Off-the-Shelf	B	—-	—-	0.494	0.417	0.192
	FineTuned	B	B	—-	0.521	0.217	0.249
	LoRA	B	—-	B	0.494	0.367	0.244
	FT(dual)	B/M	B/M	—-	**0.682**	0.612	**0.642**∗
	FT+LoRA	B/M	B	M	0.679	**0.615**∗	0.624
WavLM	Off-the-Shelf	B	—-	—-	0.516	0.418	0.184
	FineTuned	B	B	—-	0.431	0.407	0.258
	LoRA	B	—-	B	0.503	0.387	0.237
	FT(dual)	B/M	B/M	—-	0.673	0.590	0.604
	FT+LoRA	B/M	B	M	**0.680**	**0.616**∗	**0.647**∗
					BIIC-Podcast Test		
HuBERT	Off-the-Shelf	M	—-	—-	0.429	0.175	0.207
	FineTuned	M	M	—-	0.500	0.231	0.231
	LoRA	M	—-	M	0.502	**0.242**∗	0.232
	FT(dual)	M/B	M/B	—-	0.608	0.198	0.364
	FT+LoRA	M/B	M	B	**0.624**∗	0.203	**0.388**∗
WavLM	Off-the-Shelf	M	—-	—-	0.478	0.202	0.220
	FineTuned	M	M	—-	0.516	0.211	0.244
	LoRA	M	—-	M	0.523	**0.236**∗	0.252
	FT(dual)	M/B	M/B	—-	0.623	0.136	**0.405**∗
	FT+LoRA	M/B	M	B	**0.630**∗	0.191	0.370

5. Experimental Results

This section compares the performance of the explored approach in cross-lingual settings against a monolingual setup used for fine-tuning SSL models for SER. The results are reported using CCC scores obtained from test set 1 from the MSP-Podcast corpus and the standard test set provided in the BIIC-Podcast corpus. The compared approaches are listed as follows: "Off-the-Shelf" represents the off-the-shelf HuBERT and WavLM models where the original weights are frozen and only the "Head", as shown in Figure 1, is trained; "FineTuned" represents a model which has only been trained on one of the datasets by fine-tuning the "Enc." and "Head" modules, as depicted in Figure 1; "LoRA" represents a model which has only been trained on one of the datasets by fine-tuning the "LoRA" and "Head" modules while leaving the Enc. module frozen; "FT(dual)" represents a model which has been trained on both languages by fine-tuning the "Enc." and "Head" modules as depicted in Figure 1; and "FT+LoRA" represents the proposed approach described in Section 3.

5.1. In Domain and Cross-Lingual Performance

In Table 1, our "FT+LoRA" method demonstrates improved or sustained performance in in-domain scenarios, indicating that integrating a second language via LoRA preserves the model's efficacy in the original language of training. The notable exception is the performance drop observed in the WavLM model on the BIIC-Podcast corpus for valence. Although less pronounced, this trend aligns with the performance decrease seen in HuBERT. This result can be explained by the lexical cues implicitly conveyed in the SSL representation, which have been shown to be key in valence prediction [25].

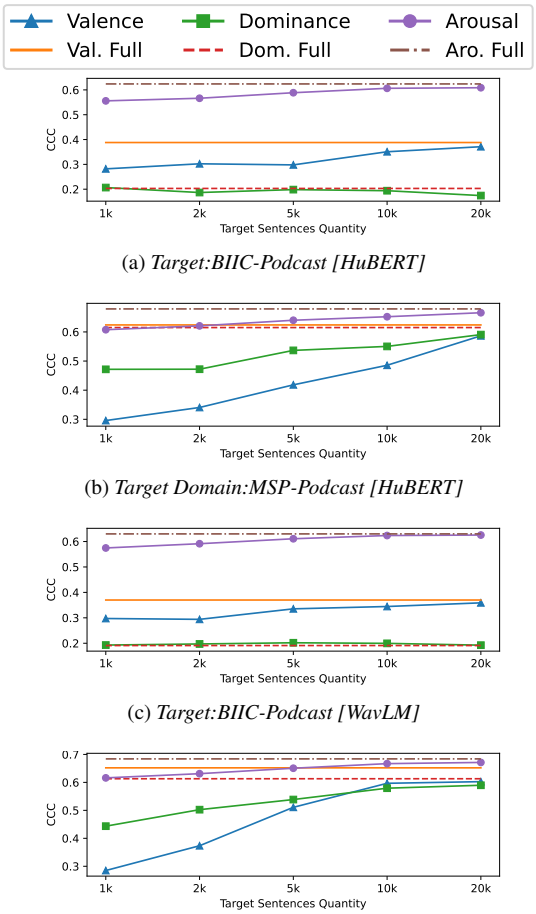
Table 2 presents the cross-lingual results for these models. As anticipated, the monolingual models "Off-the-Shelf", "LoRA", and "Finetuned" exhibit generally lower performance when applied across languages without adaptation. However, something interesting happen with dominance results: when models are trained solely on the MSP-Podcast and then directly tested on the BIIC-Podcast, dominance scores surpass those of both the adapted models and the monolingual BIIC-Podcast models. A possible explanation for this unexpected increase in dominance performance could be attributed to the nature of the emotional expression in the languages in question. English, with its specific prosodic features, might express dominance in a manner that is more universally detectable, even by models trained on non-adapted data. Another factor could be the inherent characteristics and larger nature of the MSP-Podcast training data, which may include a larger pool of expressions of dominance, leading to an enhancement in the model's ability to generalize this emotional attribute across languages.

The "FT+LoRA" approach exhibits robust cross-lingual performance, particularly with English as the target language. The upper segment of Table 2 shows that models adapted from Taiwanese Mandarin to English using "FT+LoRA" significantly surpass other methods in most configurations. In contrast, when adapting from English to Taiwanese Mandarin, as shown in the lower part of the table, the "FT+LoRA" approach still leads in arousal performance. However, it shows weaker results for valence compared to the "FT(dual)" method with WavLM. While "FT(dual)" appears to have improvements in the prediction of valence during cross-lingual testing, it falls short in dominance compared to our proposed "FT+LoRA" approach. This result suggests that "FT+LoRA" maintains a more balanced performance across emotional dimensions in a cross-lingual context.

5.2. Low-Resource Cross-Lingual Adaptation

In our experiments, we extended the experiments to scenarios with limited resources. Figure 2 presents the outcomes using our "FT+LoRA" approach, where the volume of data for *language 2* in the adaptation is constrained. We varied the quantity of data available for *language 2* for LoRA reparametrization, sampling 1K, 2K, 5K, 10K, and 20K sentences from the training set. The development and test sets remained unchanged.

Figure 2 shows the results, where the straight horizontal lines represent performance using all the sentences in the train set. The trends reveal that the amount of training data plays a key role in the performance for all datasets and models, particularly in low-resource contexts such as 1K, 2K, and 5K data samples. As anticipated, increasing the data volume gradually improves results, approaching the results obtained with the complete training set. With 20K training samples – which represent a reduction of 60% from the BIIC-Podcast and 76% from the MSP-Podcast training data – we managed to attain performance comparable to that of the complete dataset in all experimental conditions. These results highlight the efficacy of the "FT+LoRA" approach in low-resource settings, underscoring its adaptability and the critical role of data volume in cross-lingual model performance. This adaptability is particularly valuable for languages with limited available data, pointing to a promising direction for future SER research in multilingual contexts.



(a) *Target:BIIC-Podcast [HuBERT]*

(b) *Target Domain:MSP-Podcast [HuBERT]*

(c) *Target:BIIC-Podcast [WavLM]*

(d) *Target Domain:MSP-Podcast [WavLM]*

Figure 2: *Performance on multilingual settings when a limited amount of data is used for reparametrization. Horizontal straight lines across the plots represents results obtained when using the full training data.*

6. Conclusions

This study investigated the application of LoRA for multilingual SER. Our experiments demonstrated that LoRA presents a robust method for enhancing a model's capabilities to include more than one language in a SER model while preserving its original performance in monolingual contexts. Furthermore, this research underscores the efficiency of LoRA, showing that tuning a minimal number of parameters can result in substantial cross-lingual performance improvements. A promising direction for future research is to expand the scope of this approach to include more languages. This expansion could involve integrating additional LoRA reparametrization modules tailored to different languages, enriching a pre-trained SER model with a wider array of language-specific emotional information. We also aim to apply the approach to multimodal SER models [33].

7. References

[1] L. C. Matthews and B. Thakkar, "The impact of globalization on cross-cultural communication," in *Globalization*, H. Cuadra-Montiel, Ed. Rijeka: IntechOpen, 2012, ch. 13. [Online]. Available: https://doi.org/10.5772/45816

[2] K. Scherer, R. Banse, and H. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, p. 76, January 2001.

[3] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[4] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 125–131.

[5] Y. Wang and F. Tian, "Recurrent residual learning for sequence classification," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, TX, USA, November 2016, pp. 938–943.

[6] M. Bhaykar, J. Yadav, and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm," in *2013 National Conference on Communications (NCC)*. IEEE, 2013, pp. 1–5.

[7] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multilingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection," in *Speech-Prosody*, Chicago, USA, 2010.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[9] M. Jia *et al.*, "Visual prompt tuning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 709–727.

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[11] Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, and Q. Tian, "Qa-lora: Quantization-aware low-rank adaptation of large language models," 2023.

[12] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Los Alamitos, CA, USA: IEEE Computer Society, sep 2023, pp. 1–8. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ACII59096.2023.10388152

[13] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "Peft: State-of-the-art parameter-efficient fine-tuning methods," https://github.com/huggingface/peft, 2022.

[14] Y. Yu, H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. Ryu, R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y. C. Liu, T. Dinh, A. Gandhe, D. Filimonov, S. Ghosh, A. Stolcke, A. Rastrow, and I. Bulyko, "Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition," in *ASRU 2023*, 2023.

[15] 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, October-December 2019.

[16] S. Upadhyay *et al.*, "An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[17] S. Chen *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, October 2022.

[18] W.-N. Hsu *et al.*, "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.

[19] J. López, I. Cearreta, I. Fajardo, and N. Garay, "Validating a multilingual and multimodal affective database," in *Usability and Internationalization, Global and Local User Interfaces (UI-HCII 2007)*, ser. Lecture Notes in Computer Science, N. Aykin, Ed. Beijing, China: Springer Berlin Heidelberg, July 2007, vol. 4560, pp. 422–431.

[20] M. Neumann and N. Vu, "Cross-lingual and multilingual speech emotion recognition on English and French," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5769–5773.

[21] S.-w. Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5881–5885.

[22] W. Zehra, A. Javed, Z. Jalil, H. Khan, and T. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, August 2021.

[23] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6907–6911.

[24] S. Upadhyay *et al.*, "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[25] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, sep 2023.

[26] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, vol. To appear, Seoul, Republic of Korea, April 2024.

[27] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using Wav2vec 2.0 embeddings," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3400–3404.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, Dec. 2020, pp. 12 449–12 460.

[29] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6922–6926.

[30] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[31] S. wen Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[32] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[33] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, October-December 2022.